HTCondor with KRB/AFS

Setup and first experiences on the DESY interactive batch farm

Beyer Christoph & Finnern Thomas
Madison (Wisconsin), May 2018
HTCondor week

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

DESY.

The Team and the Outline

The Team

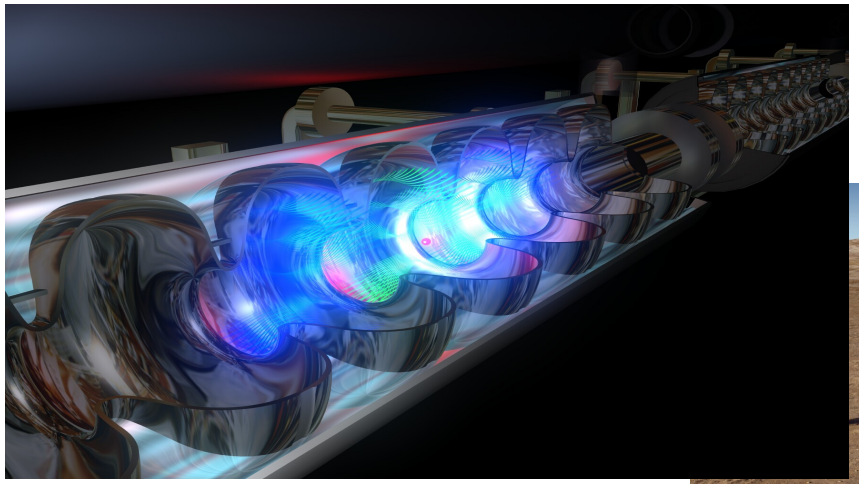
- Christoph Beyer
- Thomas Finnern (co-author)
- Martin Flemming
- Thomas Hartmann

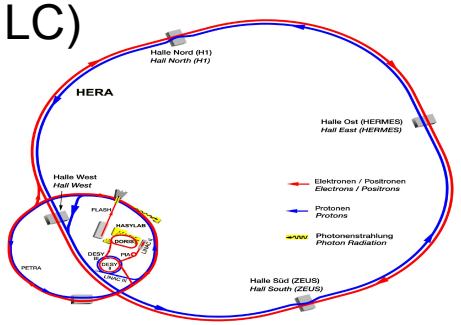
Outline of Talk

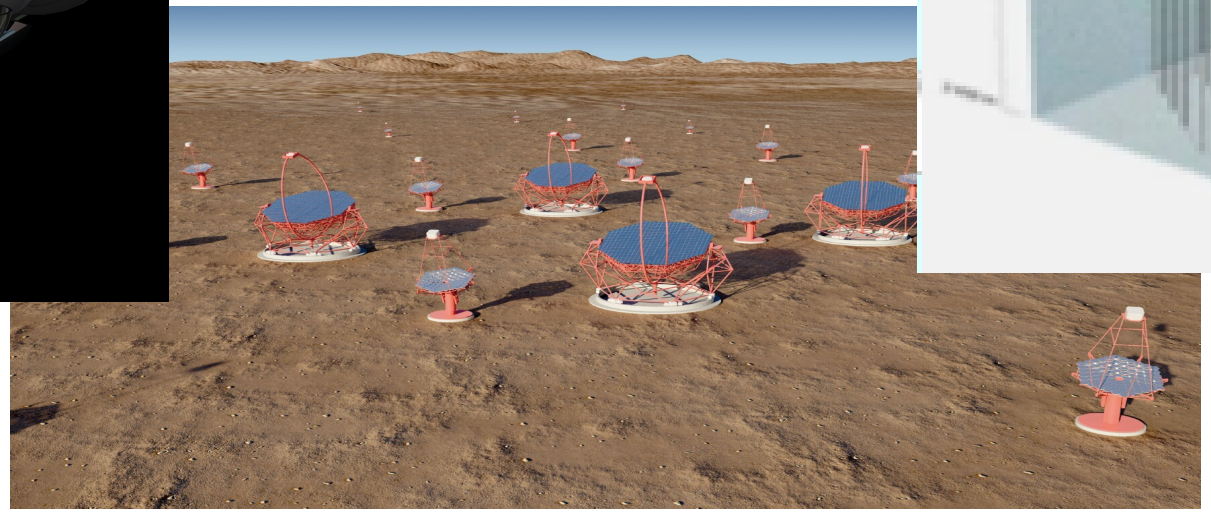
- Short introduction to DESY
- Batch facilities
- The HTCondor KRB/AFS feature
- Setup at DESY
- Migration and first usage experience
- Conclusions

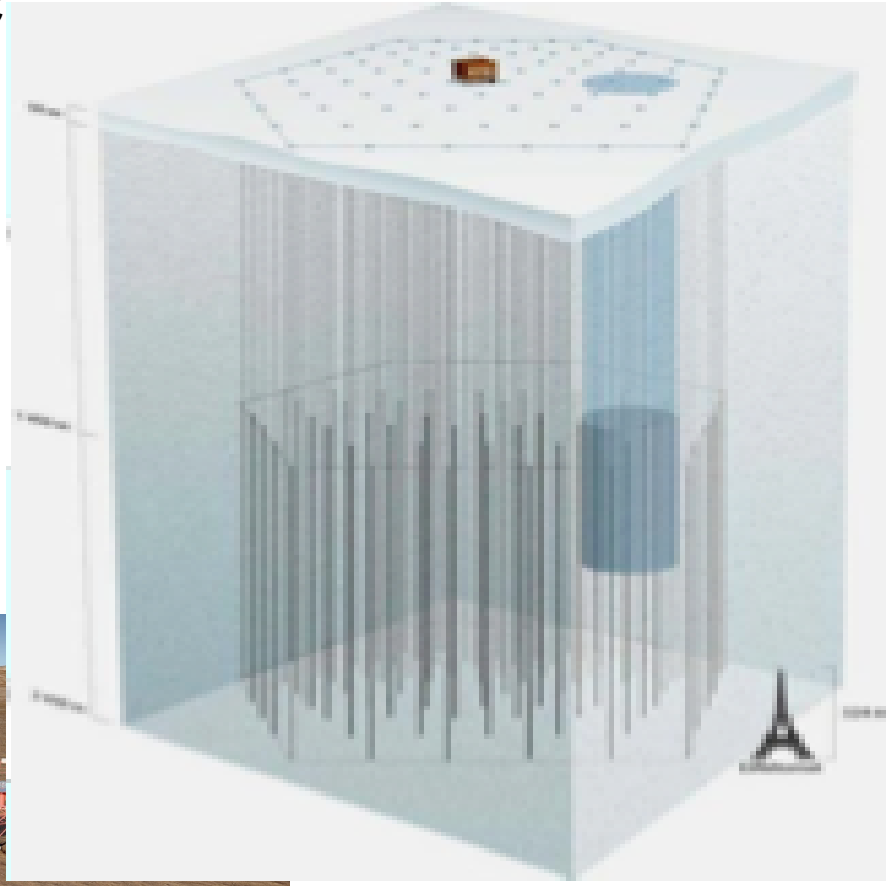
IT activities

- Operation of a 'Large Scale Facility' Tier2 and Analysis Center for LHC
- Operation of Tier0 and Tier1 for HERA, IceCube, BELLE 2
- Operation of Tier1 for Cherenkov Telescope Array (CTA)
- Preparations for a future linear collider project (ILC)



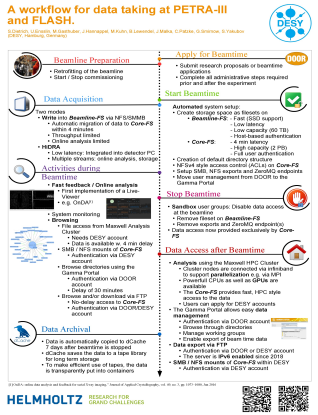


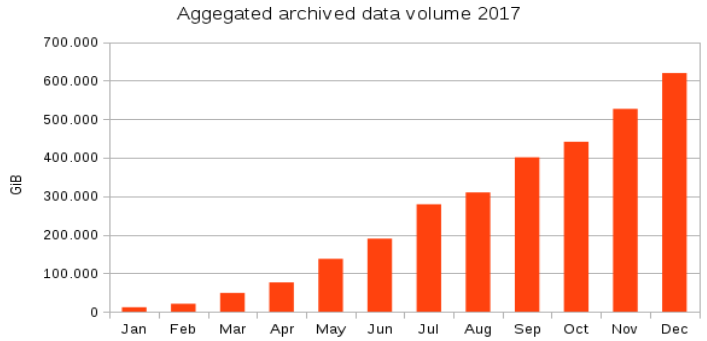


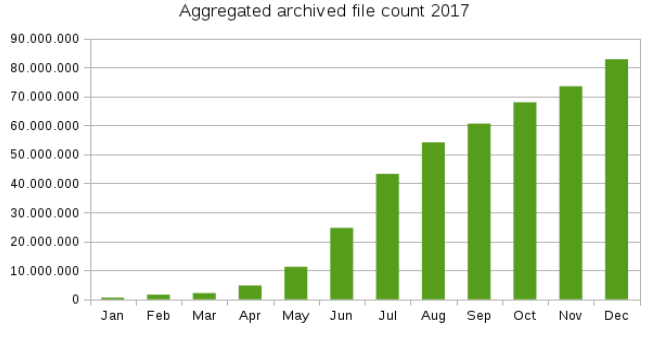


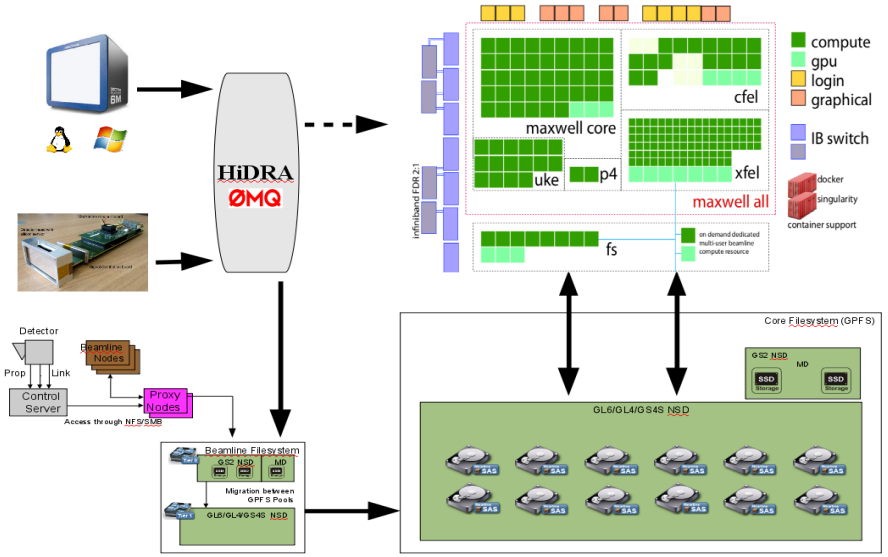
PETRA-III and FLASH

- ASAP³ data taking and management workflow is well established and running smoothly [1]
- 2 new beam lines added, more to come
- Focus is moving to data analysis
  - Maxwell HPC cluster, GPUs ...
- New data sources are being added to ASAP³
  - Detector development, on-site laser and microscope experiments, long term XFEL accelerator monitoring, ...





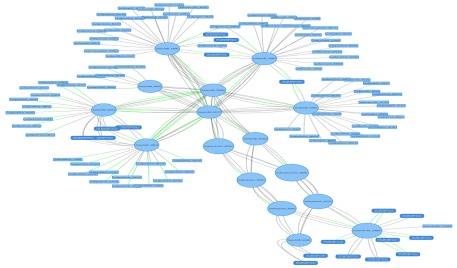


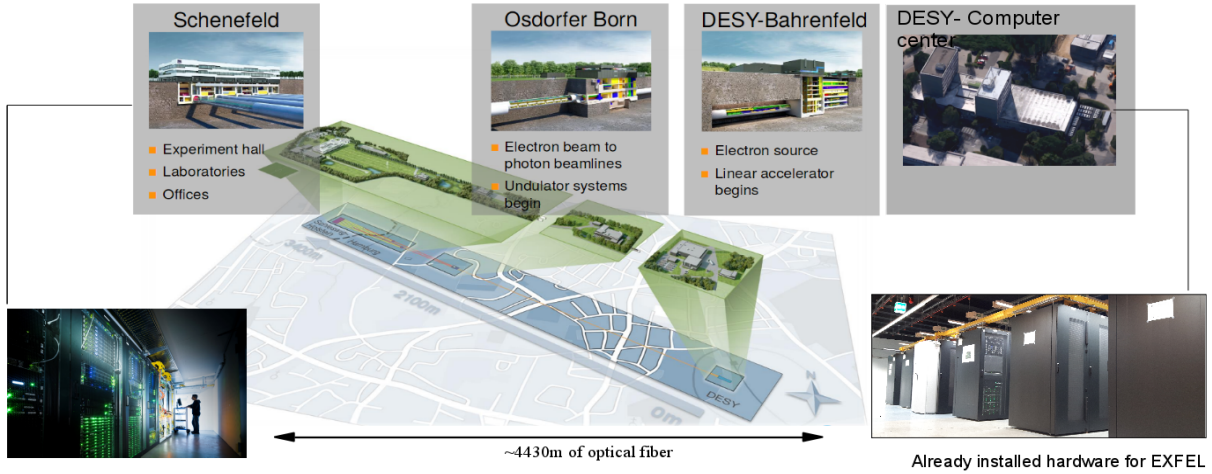


[1] Journal of Physics: Conference Series **664** (2015) 042053

European XFEL

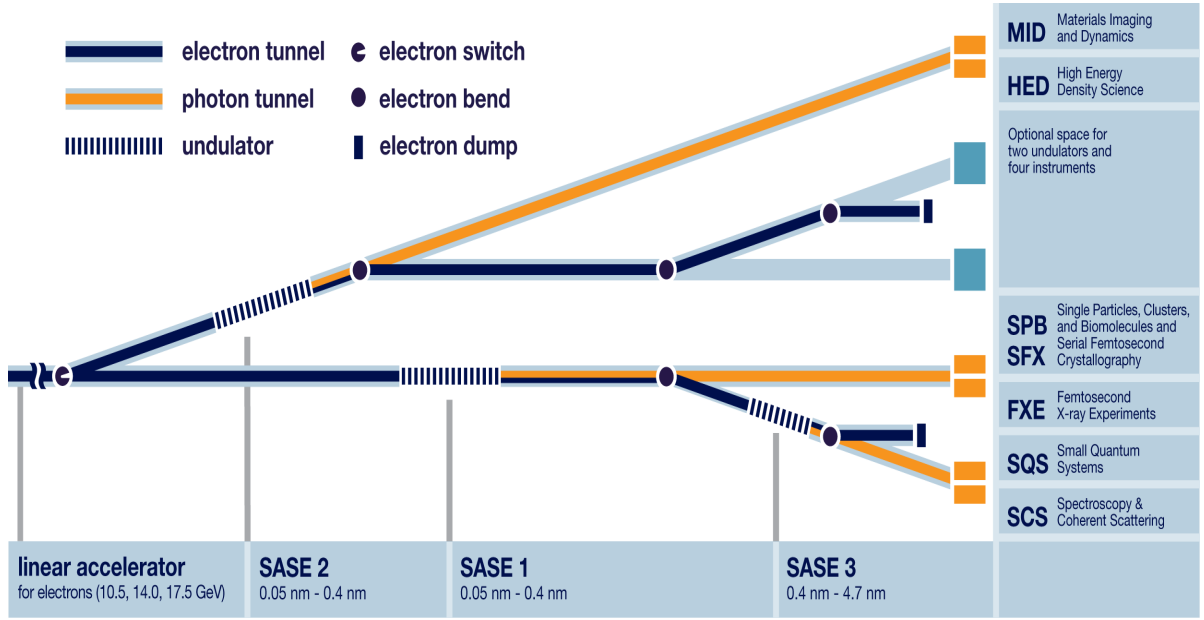
- Data taking off-campus in Schenefeld, storage & analysis on the DESY campus
- User runs in November 2017, March & May 2018
  - More continuous operation starting in summer
- About 1 PiB of data stored already, up to 50 TiB per day
- Calibration and analysis using the Maxwell HPC cluster at DESY
- One SASE active, two more starting this year
- Infiniband monitoring developed by IT





4 computer rooms in the experiment hall (a.k.a. balcony rooms)
Dedicated rack rooms for the instruments

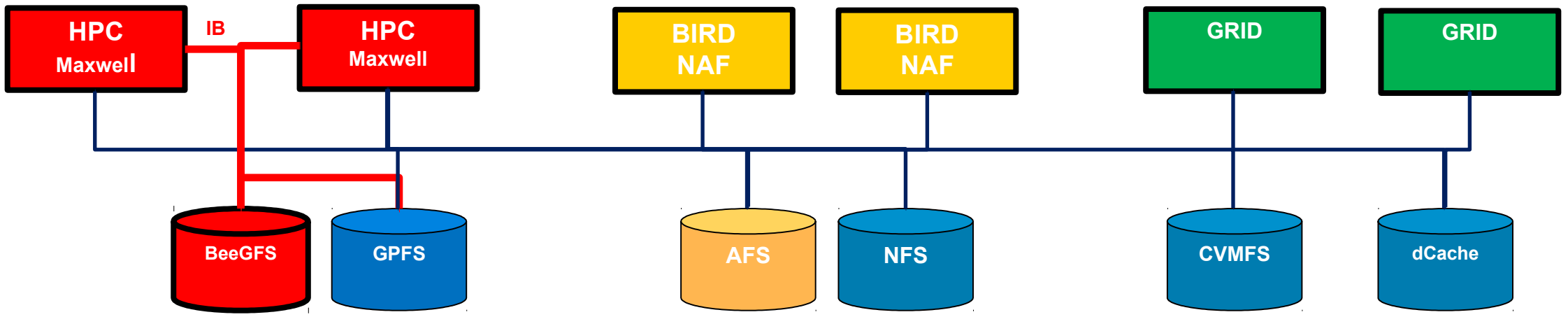
Already installed hardware for EXFEL



SASE: Self-Amplified Spontaneous Emission

Batch-clusters overview

SLURM	HTCondor (in migration)	HTCondor & ARC CE



440 nodes 24.436 cores	~ 10.000 cores (2/3 migrated)	14.500 cores

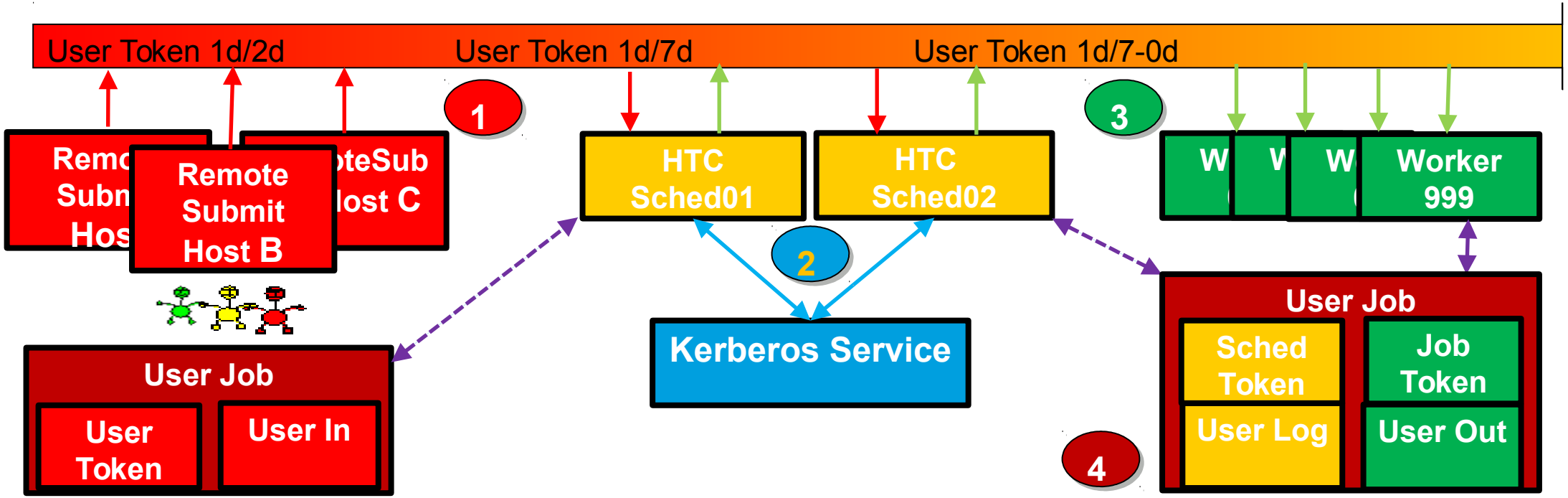
What we wanted to achieve

- Kerberos authentication at submit host
- NFS V.3 r/w access through uid/gid
- AFS as shared filesystem ($HOME for DESY user)
- Valid tokens during job run time
- Generation of AFS tokens out of Kerberos tokens
- 1 week maximal job run time
- Secure token generator on protected servers (vs. on workernodes)
- Consistently prolong current token
- Fairshare and Group quotas including surplus and balance between short/long jobs

The KRB/AFS integration in HTCondor

- KRB ticket handling looks easy at first glance but gets more complicated when you want it to be secure, reliable and easy to use
- Creating an AFS token is easy inside your job when the KRB environment is set, hence writing to afs from inside the job is easy
- Having job output, log- and errorfiles in AFS means that the HTCondor daemons on the workernode and on the scheduler need AFS tokens too
- The token should be transferred with the job for not having all available tokens on all workernodes
- You don’t want to keep credentials of users forever unless they do have running or hold jobs and will need a credential later on
- The lifetime of a standard KRB/AFS ticket is usually 24 hours and some condor jobs may run longer than that, currently guaranteed 1 week job runtime
- KRB token manipulation should be restricted to a ‘non-user-login-host’ for security reasons

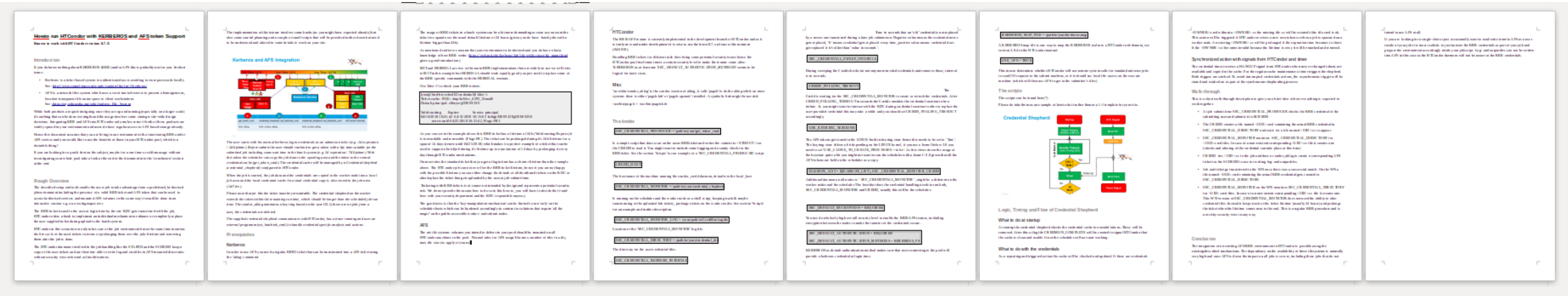
The KRB/AFS integration in HTCondor



1	2	3	4
get_token_cmd	credential_shepherd, set_batchtok_cmd	credential_shepherd, set_batchtok_cmd	(HTCondor internal)
kinit, aklog	kinit, condor_aklog	kinit, condor_aklog	

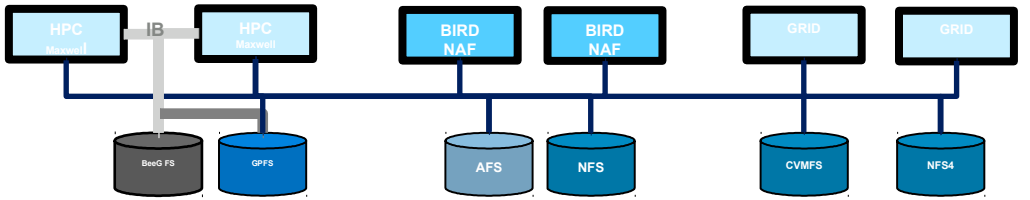
HTCondor KRB/AFS recipe

- Hopefully useful document
- Scripts for sec_credential_producer and sec_credential_monitor includes as starting point
- Token manipulation is not included
- sec_credential_producer and sec_credential_monitor will become regular condor code
- Support and documentation with 8.8



More detailed setup at DESY

- 10 Remote Submit Hosts
  - Kerberos authentication
  - ‘Project’ specific default settings
  - No depencies from/to running jobs
  - Desirable for OS administration
  - Token manipulator on SCHED
- 2 Scheduler
  - Switch scheduler for restart/service of the OS
  - Token manipulation on no-user-login host
  - Policy settings on scheduler (transforms)
  - Check ‘projects’ against registry
- 500 Worker Nodes
  - Common node setup (puppet)
  - Currently 300+ in Pilot
- 1 “CONDOR_HOST”
  - Collector and Negotiator
  - Quota/Fairshare configuration
  - Should become failsafe/clustered in the future



User Registry Integration

CLASSAD_USER_MAPFILE_Projects

- Set adequate project on group submit host
- User may switch to another project in submit file or commandline
- Project defaults to primary registry group
- Resulting project will be checked against registry on scheduler
- Resulting project will define fairshare/quota group
- Resulting project will be set on worker as primary group
- Jobs with invalid project do not run (go on hold with holdreason “wrong project ...”)

Job classes

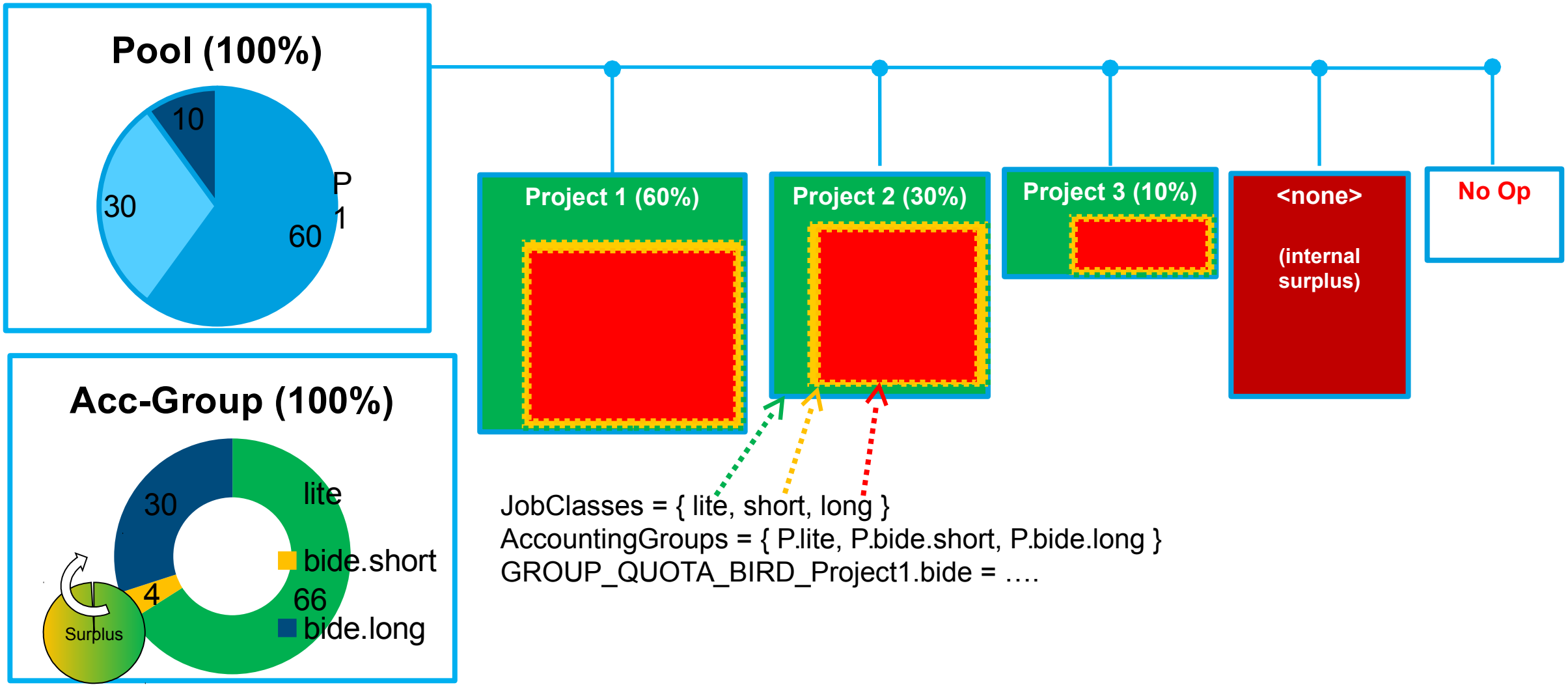
Lite/bide.short/bide.long

- Very heterogenous jobtypes (from default 3 hours to 1 week possible, no preemption other than runtime > exp. runtime)
- Hierarchical Quota with Overcommittment
  - Fairshare regulation over time
  - Partial surplus for lite jobs for fast node fill
- Job Classes and Accounting Groups
  - lite -> project.lite (default job 3 hours)
  - short -> project.bide.short (medium job 24 hours)
  - long -> project.bide.long (long job 7 x 24 hours)
  - Surplus on project.lite
- 300 % Overcommittment
  - Allows flooding the cluster with lite jobs
  - Forbids flooding the cluster with bide jobs
- Heavy usage of job_transforms

```
JOB_TRANSFORM_Y3JobClass @=end
[
eval_set_JobClass = ifThenElse(RequestMemory <= 2048 && MaxJobRetirementTime <=  3 * 3600, "lite", \
                    ifThenElse(Requestmemory <= 4096 && MaxJobRetirementTime <= 24 * 3600, "short", \
                    "long"));
eval_set_JobType = ifThenElse(RequestCpus == 1 && TotalSubmitProcs == 1, "single", \
                   ifThenElse(RequestCpus == 1 && TotalSubmitProcs != 1, "array", \
                   ifThenElse(RequestCpus != 1 && TotalSubmitProcs == 1, "multicore", \
                   "multiarray")));
]
@end
```

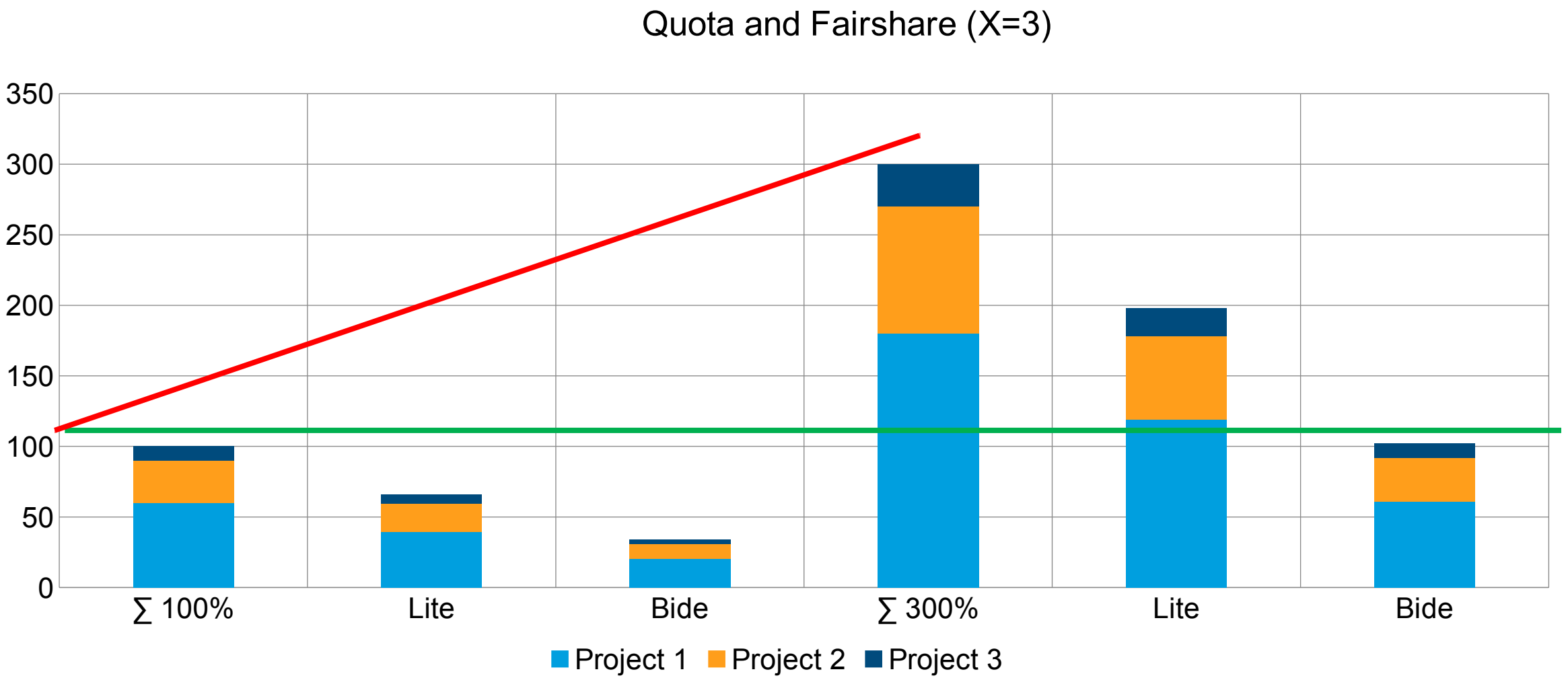
Quota and Fairshare Handling

JobtClasses



Quota and Fairshare Handling

300 % Overcommittment – make everybody equally unhappy



Quota and Fairshare Handling

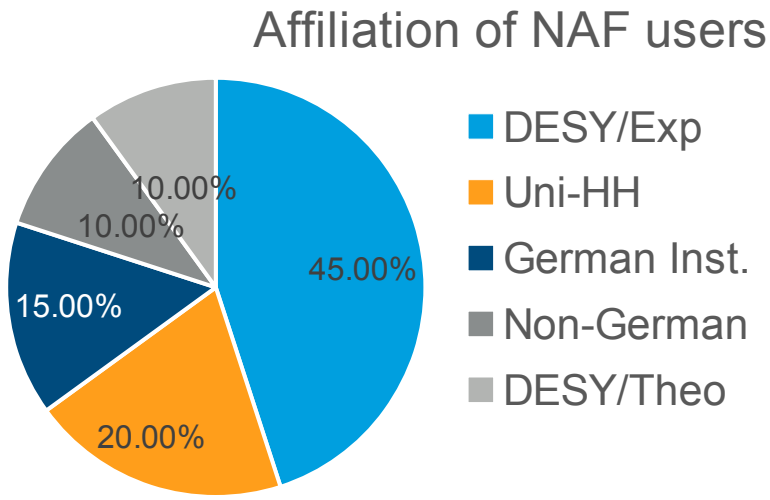
condor_userprio -most

```
[chbeyer@htc-it02]~/htcondor/testjobs% condor_userprio -most
Last Priority Update:  5/21 23:22
Group                  Config     Use    Effective   Priority   Res   Total Usage  Time Since Requested
  User Name             Quota   Surplus   Priority    Factor   In Use (wghted-hrs) Last Usage Resources
---------------------- --------- ------- ----------- --------- ------ ----------- --------- ----------
BIRD_atlas.bide         2177.42 no                    1000.00      0    547143.12    0+02:38          6
BIRD_it.lite             290.32 ByQuota               1000.00      0     35219.93    0+01:22          0
BIRD_atlas.lite         4354.84 ByQuota               1000.00      1    128108.77      <now>          2
  chenyh@desy.de                              9364.68 1000.00      1      1798.10      <now>
BIRD_cms.lite           5806.45 ByQuota               1000.00      6    284349.78      <now>          6
  zenaiev@desy.de                              762.38 1000.00      2        84.99      <now>
  stadie@desy.de                              8995.81 1000.00      4     25183.47      <now>
BIRD_ilc.bide            580.65 no                    1000.00     20     82339.02      <now>          0
BIRD_other.bide          145.16 no                    1000.00    144     96011.54      <now>       1100
  finnern@desy.de                             97079.26 1000.00    144     96011.54      <now>
BIRD_cms.bide           2903.23 no                    1000.00   2904   2202550.00      <now>      13135
  estevezl@desy.de                            75121.14 1000.00     99     26338.34      <now>
  zenaiev@desy.de                            367033.44 1000.00      1    643496.75    0+00:09
  jung@desy.de                               421575.19 1000.00   1085    572131.19      <now>
  asaibel@desy.de                           1049057.00 1000.00   1720    178231.39      <now>
---------------------- --------- ------- ----------- --------- ------ ----------- --------- ----------
Number of users: 8                ByQuota                        3056   1543275.75    0+23:59
```

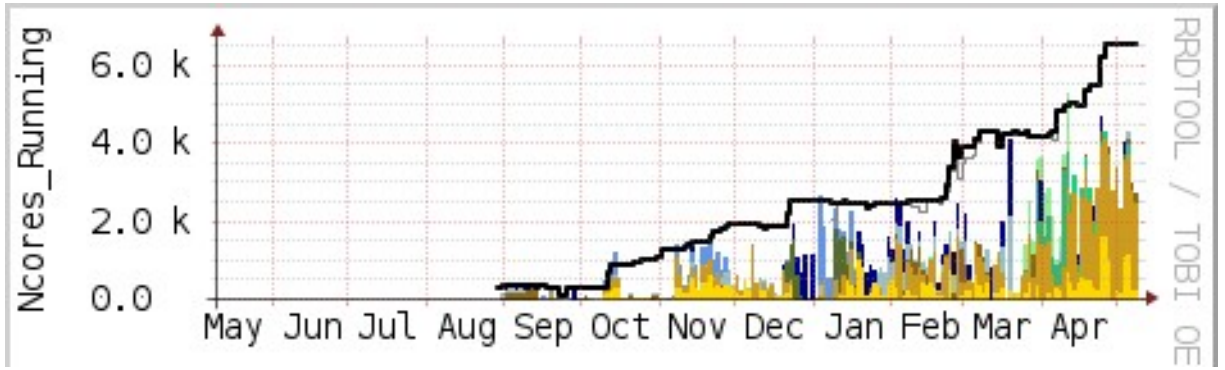
BIRD / NAF

Migration from SGE to HTCondor in progress

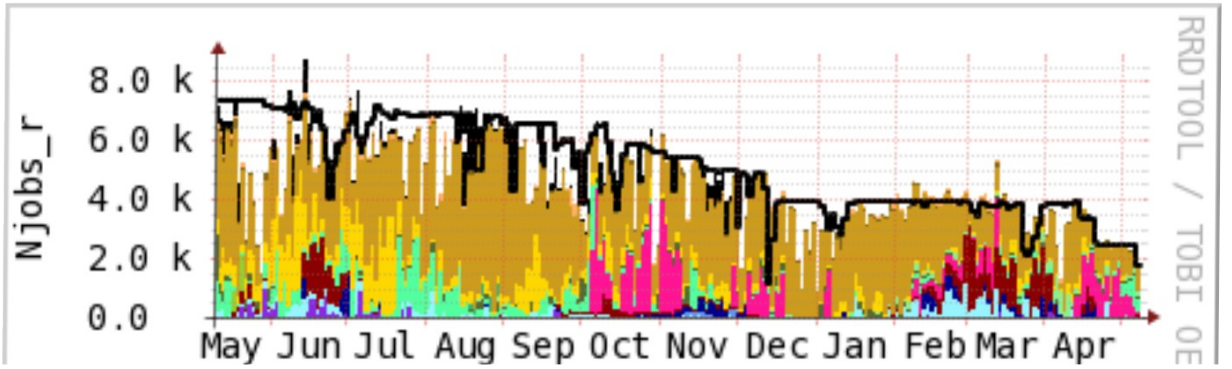
- Roughly a year between first ideas and actual start of migration
- AFS/KRB feature needed some more further attention from the HTCondor team than expected at first
- Distributed FS mean dependencies of course, even more with HTCondor than with SGE
- Teaching the users is essential



HTCondor pilot



SGE



Conclusions

Given our experience so far

- KRB/AFS feature very usable
- Depending on shared filesystems is a thing
- Token manipulation is evil and should be done with more ‘KRB-internal’ tools
- User education is essential
- Support by the HTCondor team is very helpful !
- Infrastructure maybe subject to reconsideration, especially the ‘single’ scheduler
- Need to put more effort in monitoring
- Job return codes need to be investigated
- Job_wrapper should be avoided somehow
- Need more intelligence for the STARTD_CRON
- Token cleanup on workernodes needs inspection
- SEC_CREDENTIAL_MONITOR will be replaced with 8.8 by a python version
- Sanity checks for tokens
- Check for ways to unite GRID and BIRD pools
  - Both pools very busy
  - Very different usage patterns
  - Dependency to shared FS for GRID user
  - Common tools/monitoring for both pools in use
  - Much potential in the opportunistic usage of HPC-on-campus ressources

Thank you

Contact

DESY. Deutsches
Elektronen-Synchrotron

www.desy.de

Beyer Christoph
IT-Department
Christoph.beyer@desy.de
++49 40 8998 2317